A bayesian solution to the Behrens-Fisher problem

Contact: jean-marc.freyermuth@univ-amu.fr

Master DS, Academic year, 2024-2025

Summary

- multiparameter Bayesian models; case of a normal distribution with unknown mean and variance,
- comparison of two normal means: the Behrens-Fisher problem,
- a solution using the Gibbs sampler,
- “Modern” Behrens-Fisher problems.

Multiparameter models

Generally many parameters are involved in a statistical model.

Some of them are not of interest: **nuisance parameters.**

2 parameters case: consider $\boldsymbol{\theta} = (\theta_1, \theta_2)$, $\pi(\boldsymbol{\theta})$ is the associated **joint prior** distribution. We compute the **joint posterior** using the bayes rule:

$$\pi(\boldsymbol{\theta}|x) = \pi(\theta_1, \theta_2|x) \propto \pi(\theta_1, \theta_2)p(x|\theta_1, \theta_2).$$

As a consequence, if one wants to make inference on θ_1, we integrate out θ_2

$$\pi(\theta_1|x) = \int_{\Theta_2} \pi(\theta_1, \theta_2|x)d\theta_2$$

Terminology:

- **posterior marginal** of θ_1: $\pi(\theta_1|x)$.
- **posterior conditional** of θ_1 given θ_2: $\pi(\theta_1|\theta_2, x)$.

A remark on invariant prior specification

Location

If the parameter of interest is a location parameter θ, i.e., $x|\theta \sim p(x-\theta)$. A proper non informative prior has to be invariant w.r.t translations, i.e.,

$$\pi(\theta - \theta_0) = \pi(\theta), \ \forall \theta_0.$$

Hence $\pi(\theta) = constant$.

remark:

- if the parameter space is unbounded, this prior is not a p.d.f, this is an **improper prior**. This is fine as long as the posterior is proper.
- sufficient condition for obtaining a proper posterior is that the prior predictive distribution is finite for any x.

A remark on invariant prior specification

Scale

If the parameter of interest is a scale parameter, i.e., $x|\theta \sim \frac{1}{\theta}p(\frac{x}{\theta})$. The prior has to be scale invariant.

$$\pi(\theta) = \frac{1}{c}\pi\left(\frac{\theta}{c}\right), \ \forall c > 0.$$

Hence, we choose $\pi(\theta) \propto \frac{1}{\theta}$.

Normal model with unknown mean and variance

Reminder: important distributions.

- the Gamma distribution with parameters (λ, α)

$$
\begin{aligned}
x &\sim \Gamma(\lambda, \alpha) \\
p(x) &= \frac{\lambda^{\alpha} x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)},\ x > 0, \lambda, \alpha > 0.
\end{aligned}
$$

- the Inverse Gamma distribution with parameters (λ, α)

Let $y = 1/x$ where $x \sim \Gamma(\lambda, \alpha)$, then

$$
\begin{aligned}
y &\sim \Gamma^{-1}(\lambda, \alpha) \\
p(y) &= \frac{\lambda^{\alpha} y^{-(\alpha+1)} e^{-\lambda/y}}{\Gamma(\alpha)},\ y > 0, \lambda, \alpha > 0.
\end{aligned}
$$

We observe $(x_1, \ldots, x_n)|\mu, \sigma^2 \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$. Consider the following prior distribution:

$$
\pi(\mu, \sigma^2) = \pi(\mu)\pi(\sigma^2) \propto \frac{1}{\sigma^2}.
$$

Normal model with unknown mean and variance

The joint posterior is easily obtained

$$\pi(\mu,\sigma^2|\bar{x}) \propto \sigma^{-n-2} \exp\left\{-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{x}-\mu)^2]\right\},$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

We can write

$$\pi(\mu,\sigma^2|\bar{x}) = \pi(\mu|\sigma^2,\bar{x})\pi(\sigma^2|\bar{x}).$$

The conditional posterior of μ given σ^2 is

$$\begin{aligned}&\pi(\mu|\sigma^2,\bar{x}) \propto \exp\left\{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right\}\\ &\mu|\sigma^2,\bar{x} \sim \mathcal{N}(\bar{x},\sigma^2/n).\end{aligned}$$

Normal model with unknown mean and variance

Marginal posterior of σ^2

$$\begin{aligned}
\pi(\sigma^2|\bar{x}) &= \int_{-\infty}^{\infty} \pi(\mu, \sigma^2|\bar{x}) d\mu \\
&\propto \sigma^{-n-2} \exp\left\{-\frac{1}{2\sigma^2}(n-1)s^2\right\} \sqrt{2\pi\sigma^2/n} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left\{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right\} d\mu \\
&\propto (\sigma^2)^{-[(n-1)/2+1]} \exp\left\{-\frac{1}{2\sigma^2}(n-1)s^2\right\}.
\end{aligned}$$

I.e., $\sigma^2|\bar{x} \sim \Gamma^{-1}\left((n-1)s^2/2, (n-1)/2\right)$.

Normal model with unknown mean and variance

Marginal posterior of μ

$$\begin{aligned}\pi(\mu|\bar{x}) &= \int_0^\infty \pi(\mu,\sigma^2|\bar{x})d\sigma^2 \\ &= \int_0^\infty \pi(\mu|\sigma^2,\bar{x})\pi(\sigma^2|\bar{x})d\sigma^2 \\ &\propto \left(1+\frac{n(\mu-\bar{x})^2}{(n-1)s^2}\right)^{-n/2}.\end{aligned}$$

which is a generalized student (mixture of normals for different values of inverse-gamma distributed variances).

$$\mu|\bar{x} \sim t\left(\bar{x}, \frac{s}{\sqrt{n}}, n-1\right).$$

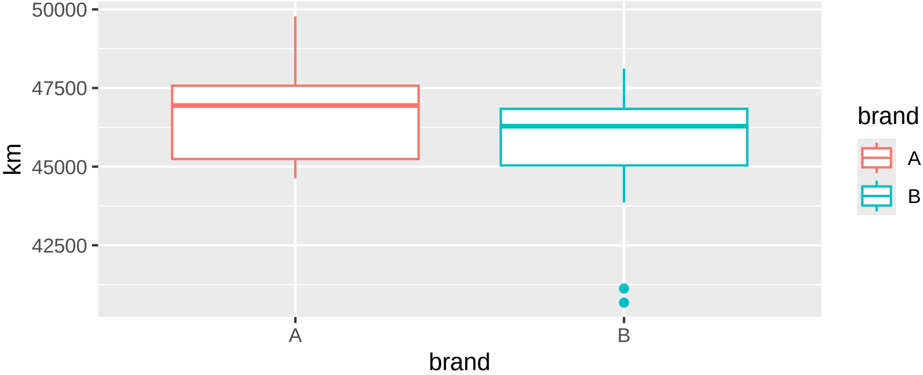
Comparing means of two samples (1)

(Example inspired from Simar, L. (2002)). A firm wants to compare the quality of two differents brand of tires w.r.t their lifespan (number of kilometers to drive before tires are too damaged). This lifespan will vary from on tire to the other because of fluctuations in the production process (considering condition of experiments are controlled). We believe that a normal distribution can modelled these fluctuations.

```r
head(dat)
```

```
##          km brand
## 1 45072.57     A
## 2 46789.13     A
## 3 44629.33     A
## 4 49780.34     A
## 5 47098.21     A
## 6 44661.46     A
```

Comparing means of two samples (2)



Comparing means of two samples (3)

Modelling hypothesis: two **independent** samples of sizes n_1, n_2.

$$x_{1i}|\mu_1, \sigma_1^2 \overset{i.i.d}{\sim} \mathcal{N}(\mu_1, \sigma_1^2),\ 1 \leq i \leq n_1.$$

$$x_{2j}|\mu_2, \sigma_2^2 \overset{i.i.d}{\sim} \mathcal{N}(\mu_2, \sigma_2^2),\ 1 \leq j \leq n_2.$$

Comparing means of two samples (4)

Question: compare the mean of two normal populations based on two independent random samples of resistance measures from tires produced by two companies.

$$\delta = \mu_1 - \mu_2.$$

Consider 3 situations:

- known variances,
- unknown but equal variances,
- unknown and unequal variances (this problem is known as the **Berhens-Fisher problem**).

Comparing means of two samples: with known variances (1)

We decide to represent the prior information on the two means μ_1, μ_2 by two independent normal distributions.

$$\mu_k \sim \mathcal{N}(m_{k0}, \eta_{k0}^{-1}),\ k = \{1, 2\},\ \mu_1 \perp \mu_2,$$

where η_{k0} represents the **precision** $(\eta_{k0} = 1/\sigma_{k0}^2)$, $k = 1, 2$. We know that $\bar{x}_1$ and $\bar{x}_2$ are **sufficient** statistics. Then,

$$\mu_k | \bar{x}_k \sim \mathcal{N}(m_k^*, v_k^*).$$

where

$$m_k^* = \frac{m_{k0}\eta_{k0} + \bar{x}_k \eta_{n_k}}{\eta_{k0} + \eta_{n_k}},$$
$$v_k^* = (\eta_{k0} + \eta_{n_k})^{-1}.$$

with $\eta_{n_k}^{-1} = \frac{\sigma_k^2}{n_k}$.

Comparing means of two samples: with known variances (2)

Given the independence between the two samples, using properties of normals random variables, the posterior distribution of the parameter of interest δ is:

$$\delta | \bar{x}_1, \bar{x}_2 \sim \mathcal{N}(m_1^* - m_2^*, v_1^* + v_2^*).$$

Remark

- a) if $\eta_{k0} \to 0,\ k = 1, 2$. Non informative prior, this posterior becomes

$$\delta | \bar{x}_1, \bar{x}_2 \sim \mathcal{N}\left(\bar{x}_1 - \bar{x}_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

- b) we have analogous result to the classical frequentist paradigm.

Here we are interested by posterior probability of $\delta > 0$ to decide if tires from the one company are better than the ones of the other.

Comparing means of two samples: with unknown but equal variances (1)

we will see that the *posterior law of δ is still analytically tractable.*

Suppose $\sigma_1^2 = \sigma_2^2 = \sigma^2$

The parameters of the model are (μ_1, μ_2, σ^2) and $(\bar{x}_1, \bar{x}_2, s^2)$ are sufficient statistics for (μ_1, μ_2, σ^2), s^2 is the standard unbiased estimator of σ^2 (pooled sample variance estimator).

$$s^2 = v^{-1}(v_1 s_1^2 + v_2 s_2^2).$$

with

$$\begin{aligned} v_k &= n_k - 1, k = 1, 2, \\ s_k^2 &= v_k^{-1} \sum_{i=1}^{n_k} (x_{ki} - \bar{x}_k)^2, \ \ k = 1, 2, \\ v &= v_1 + v_2 = n_1 + n_2 - 2. \end{aligned}$$

Comparing means of two samples: with unknown but equal variances (2)

For simplicity, set a **non informative prior**: we take independent non informative prior on (μ_1, μ_2, σ^2).

$$\pi(\mu_1, \mu_2, \sigma^2) \propto \frac{1}{\sigma^2}.$$

The likelihood function is obtained taking into account the sampling independence

$$\begin{aligned} p(\bar{x}_1, \bar{x}_2, s^2 | \mu_1, \mu_2, \sigma^2) &= p(\bar{x}_1 | s^2, \mu_1, \mu_2, \sigma^2) p(\bar{x}_2 | s^2, \mu_1, \mu_2, \sigma^2) p(s^2 | \mu_1, \mu_2, \sigma^2) \\ &= p(\bar{x}_1 | \mu_1, \sigma^2) p(\bar{x}_2 | \mu_2, \sigma^2) p(s^2 | \sigma^2). \end{aligned}$$

where

$$\begin{cases} \bar{x}_k | \mu_k, \sigma^2 \sim \mathcal{N}\left(\mu_k, \dfrac{\sigma^2}{n_k}\right), \ k = 1, 2. \\ s^2 | \sigma^2 \sim \Gamma\left(\dfrac{v}{2\sigma^2}, \dfrac{v}{2}\right). \end{cases}$$

Comparing means of two samples: with unknown but equal variances (3)

Then we compute the posterior of (μ_1, μ_2, σ^2) using

$$\pi(\mu_1, \mu_2, \sigma^2 | \bar{x}_1, \bar{x}_2, s^2) \propto p(\bar{x}_1, \bar{x}_2, s^2 | \mu_1, \mu_2, \sigma^2) \pi(\mu_1, \mu_2, \sigma^2).$$

which is

$$\pi(\mu_1, \mu_2, \sigma^2 | \bar{x}_1, \bar{x}_2, s^2) = \frac{1}{\sqrt{2\pi\sigma^2/n_1}} \exp\left(-\frac{(\bar{x}_1 - \mu_1)^2}{2\sigma^2/n_1}\right) \frac{1}{\sqrt{2\pi\sigma^2/n_2}} \exp\left(-\frac{(\bar{x}_2 - \mu_2)^2}{2\sigma_2^2/n_2}\right)$$
$$\times \left(\frac{v}{2\sigma^2}\right)^{v/2} \frac{(s^2)^{(v/2-1)} \exp\left\{-\frac{v}{2\sigma^2} s^2\right\}}{\Gamma(v/2)} \frac{1}{\sigma^2}.$$

Comparing means of two samples: with unknown but equal variances (4)

We factorize the posterior as follows:

$$\pi(\mu_1, \mu_2, \sigma^2 | \bar{x}_1, \bar{x}_2, s^2) = \pi(\mu_1 | \bar{x}_1, \sigma^2)\pi(\mu_2 | \bar{x}_2, \sigma^2)\pi(\sigma^2 | s^2),$$

where

$$\begin{cases} \mu_k | \sigma^2, \bar{x}_k \sim \mathcal{N}\left(\bar{x}_k, \dfrac{\sigma^2}{n_k}\right). \\ \sigma^2 | s^2 \sim (vs^2)\chi_v^{-2}. \end{cases}$$

Comparing means of two samples: with unknown but equal variances (5)

The posterior law of σ^2 is proportional to a χ_v^{-2}, it is therefore an inverse gamma

$$\begin{cases} \sigma^2 | s^2 \sim \Gamma^{-1}\left(\frac{vs^2}{2}, \frac{v}{2}\right) \\ \pi(\sigma^2|s^2) = \frac{1}{\Gamma(v/2)}\left(\frac{vs^2}{2}\right)^{v/2} (\sigma^2)^{-[v/2+1]} \exp\left(-\frac{vs^2}{2\sigma^2}\right) \end{cases}$$

In particular we have

$$\mathbb{E}\left(\sigma^2|s^2\right) = s^2 \frac{v}{v-2}$$

The conditional posterior distribution of δ given σ^2

$$\delta|\sigma^2, \bar{x}_1, \bar{x}_2 \sim \mathcal{N}\left(\bar{x}_1 - \bar{x}_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

Comparing means of two samples: with unknown but equal variances (6)

We know compute the marginal posterior distribution for δ (we need to get rid off σ^2)

$$
\begin{aligned}
\pi(\delta|\bar{x}_1, \bar{x}_2, s^2) &= \int_0^\infty \pi(\delta, \sigma^2|\bar{x}_1, \bar{x}_2, \sigma^2) d\sigma^2 \\
&= \int_0^\infty \pi(\delta|\sigma^2, \bar{x}_1, \bar{x}_2, s^2)\pi(\sigma^2|\bar{x}_1, \bar{x}_2, s^2) d\sigma^2 \\
&= \int_0^\infty \pi(\delta|\sigma^2, \bar{x}_1, \bar{x}_2)\pi(\sigma^2|s^2) d\sigma^2.
\end{aligned}
$$

Comparing means of two samples: with unknown but equal variances (7)

We need now solve this integral. For that we need the properties of the Gamma function from which we get:

$$\pi(\delta|\bar{x}_1, \bar{x}_2, s^2) = \frac{\left(vs^2(1/n_1 + 1/n_2)\right)^{-1/2}}{B(1/2, v/2)} \left(1 + \frac{[\delta - (\bar{x}_1 - \bar{x}_2)]^2}{vs^2[1/n_1 + 1/n_2]}\right)^{-(v+1)/2}.$$

which is the density of a generalized student

$$\delta|\bar{x}_1, \bar{x}_2, s^2 \sim t(\bar{x}_1 - \bar{x}_2, s^2(1/n_1 + 1/n_2), v).$$

where $v = n_1 + n_2 - 2$. In particular we have

$$\mathbb{E}\left[\delta|\bar{x}_1, \bar{x}_2, s^2\right] = \bar{x}_1 - \bar{x}_2.$$
$$V\left[\delta|\bar{x}_1, \bar{x}_2, s^2\right] = s^2\left(1/n_1 + 1/n_2\right)\frac{v}{v-2}.$$

We therefore find analogous results to classical ones, but here we can compute the posterior probabilities that δ take any values in a set $A \subset \mathbb{R}$. $P\left[\delta \in A|\bar{x}_1, \bar{x}_2, s^2\right]$

Comparing means of two samples: with unknown but equal variances (8)

Numerical application: Compute the probability that $\delta > 0$.

```r
library(LaplacesDemon, quietly = TRUE)
M =  10000
diff_mean = x1.bar - x2.bar
nu1  = n1-1
nu2 = n2-1
nu = nu1+ nu2 - 2
pooled_var = (s1^2*nu1 + s2^2*nu2)/nu
y  = rst(n = M, mu = diff_mean, sigma =
           sqrt(pooled_var*(1/n1+1/n2)), nu = nu)
```

```
## posterior probability of 0.9124  that the average lifetime
##     of tires of brand A is larger than that of brand B
```

Comparing means of two samples: with unknown variances (1)

In the frequentist framework, there are **no exact solution** for finite sample size, only asymptotic approximations are obtained (this is the Berhens-Fisher problem). In the bayesian framework we can get an easy solution. Here this is a four parameter model but the parameter of interest is still $\delta = \mu_1 - \mu_2$.

We keep the same approach with independent and non informative prior over all the parameters

$$\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \propto \frac{1}{\sigma_1^2}\frac{1}{\sigma_2^2}.$$

For the likelihood function

$$p(\bar{x}_1, \bar{x}_2, s_1^2, s_2^2 | \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \propto p(\bar{x}_1 | \mu_1, \sigma_1^2) p(\bar{x}_2 | \mu_2, \sigma_2^2) p(s_1^2 | \sigma_1^2) p(s_2^2 | \sigma_2^2),$$

where

$$\bar{x}_k | \mu_k, \sigma_k^2 \sim \mathcal{N}(\mu_k, \sigma_k / n_k),\ \ k = 1, 2,$$
$$s_k^2 | \sigma_k^2 \sim \frac{\sigma_k^2}{v_k} \chi^2_{v_k} \sim \Gamma(\frac{v_k}{2\sigma_k^2}, \frac{v_k}{2}), k = 1, 2.$$

Comparing means of two samples: with unknown variances (2)

Similarly one can show that the posterior factorizes as follows

$$\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|\bar{x}_1, \bar{x}_2, s_1^2, s_2^2) = \pi(\mu_1|\sigma_1^2, \bar{x}_1)\pi(\mu_2|\sigma_2^2, \bar{x}_2)\pi(\sigma_1^2|s_1^2)\pi(\sigma_2^2|s_2^2),$$

where

$$\mu_k|\bar{x}_k, \sigma_k^2 \sim \mathcal{N}(\bar{x}_k, \sigma_k/n_k).$$
$$\sigma_k^2|s_k^2 \sim \Gamma^{-1}\left(\frac{v_k}{2s_k^2}, \frac{v_k}{2}\right).$$

Comparing means of two samples: with unknown variances (3)

The joint posterior marginal distribution of (μ_1, μ_2).

$$\pi(\mu_1, \mu_2 | \bar{x}_1, \bar{x}_2, s_1^2, s_2^2) = \pi(\mu_1 | \bar{x}_1, s_1^2)\pi(\mu_2 | \bar{x}_2, s_2^2).$$

where marginal posterior of μ_k are obtained as follows:

$$\begin{aligned}\pi(\mu_k | \bar{x}_k, s_k^2) &= \int_0^\infty \pi(\mu_k | \bar{x}_k, s_k^2)\pi(\sigma_k^2 | s_k^2) d\sigma_k^2 \\ &= \int_0^\infty f_N(\mu_k | \bar{x}_k, \sigma_k^2) f_{i\gamma}\left(\sigma_k^2 | \frac{v_k s_k^2}{2}, \frac{v_k}{2}\right) d\sigma_k^2.\end{aligned}$$

By definition of the generalized student law we recognized that

$$\mu_k | \bar{x}_k, s_k^2 \sim t\left(\bar{x}_k, \frac{s_k^2}{n_k}, v_k\right).$$

Comparing means of two samples: with unknown variances (4)

The marginal a posteriori of $\delta = \mu_1 - \mu_2$.

This law is obtained by simple transformation, let us do the following change of variables $(\mu_1, \mu_2) \to (\delta, \mu_2)$. Hence

$$\pi(\delta, \mu_2 | \bar{x}_1, \bar{x}_2, s_1^2, s_2^2) = \pi_{\mu_1}(\delta + \mu_2 | \bar{x}_1, s_1^2) \pi_{\mu_2}(\mu_2 | \bar{x}_2, s_2^2).$$

Then we need to integrate over μ_2:

$$\begin{aligned}
\pi(\delta | \mu_1, \mu_2 \bar{x}_1, \bar{x}_2, s_1^2, s_2^2) &= \int_{-\infty}^{\infty} \pi(\delta, \mu_2 | \bar{x}_1, \bar{x}_2, s_1^2, s_2^2) d\mu_2 \\
&= \int_{-\infty}^{\infty} \frac{[v_1 s_1^2 / n_1]^{-1/2}}{B(1/2, v_1/2)} \left(1 + \frac{n_1 (\delta + \mu_2 - \bar{x}_1)^2}{v_1 s_1}\right)^{-(v_1+1)/2} \\
&\times \frac{[v_2 s_2^2 / n_2]^{-1/2}}{B(1/2, v_2/2)} \left(1 + \frac{n_2 (\mu_2 - \bar{x}_2)^2}{v_2 s_2}\right)^{-(v_1+1)/2} d\mu_2.
\end{aligned}$$

Marginal a posteriori of the ratio of variances

We can compare the variances using the following ratio:

$$\gamma = \frac{\sigma_2^2}{\sigma_1^2}.$$

We need to get the posterior of this new parameter. From the conditional distribution and from properties of

$$\frac{\sigma_k^2}{(v_k s_k^2)}|s_k^2 \sim \chi_{v_k}^{-2}, \ k = 1, 2.$$

$$\frac{(v_k s_k^2)}{\sigma_k^2}|s_k^2 \sim \chi_{v_k}^2.$$

Marginal a posteriori of the ratio of variances

Given the independence of the two samples and between the prior information on σ_1^2 and σ_2^2, these χ^2 are independent. Then we usually have

$$\gamma \frac{s_1^2}{s_2^2} | s_1^2, s_2^2 \sim F_{\eta_1, \eta_2}.$$

$$\mathbb{E}(\gamma | s_1^2, s_2^2) = \frac{s_1^2}{s_2^2} \frac{\eta_2}{\eta_2 - 2}.$$

$$\mathbb{V}(\gamma | s_1^2, s_2^2) = \left(\frac{s_1^2}{s_2^2} \right)^2 \frac{2\eta_2^2(\eta_1 + \eta_2 - 2)}{\eta_1(\eta_2 - 2)^2(\eta_2 - 4)}, \ \eta_2 > 4.$$

$(1 - \alpha)100\%$ Credible interval for γ

$$P\left(\frac{s_1^2}{s_2^2} F_{\frac{\alpha}{2}, \eta_1, \eta_2} \leq \gamma \leq \frac{s_1^2}{s_2^2} F_{1-\frac{\alpha}{2}, \eta_1, \eta_2} | s_1^2, s_2^2 \right) = 1 - \alpha.$$

Gibbs sampler in a nutshell

- Often the posterior are P-variate distributions that do not correspond to any known distribution.

We would like to

- obtain posterior marginal distributions,
- compute their properties such as their means or a tail-areas.

If we could generate a sample of size M from the joint posterior

$$\left\{ \left(\theta_1^{(m)}, \ldots, \theta_P^{(m)} \right); 1 \leq m \leq M \right\},$$

then the $\left\{ \theta_1^{(m)}; 1 \leq m \leq M \right\}$ is a sample from the marginal posterior $\pi(\theta_1|x)$.

Using the **Monte Carlo Principle** we can compute quantities of interest since

$$\mathbb{E}\left[g(\theta_1)\right] = \int g(\theta_1)\pi(\theta_1|x)d\theta_1 \approx \frac{1}{M}\sum_{m=1}^{M} g\left(\theta_1^{(m)}\right).$$

Gibbs sampler in a nutshell

Our aim is to draw random samples from the posterior $\pi(\boldsymbol{\theta}|x)$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)' \in \mathbb{R}^P$.

- It could be difficult to obtain independent sample from the posterior but easier to find a way to generate a Markov chain which stationary distribution is our target posterior.
- *If we are in the specific situation* where we can draw samples from **all the full conditional posterior** distributions; i.e., $\pi(\theta_p|\theta_1, \ldots, \theta_{p-1}, \ldots, \theta_{p+1}, \ldots, \theta_P, x)$ **then** we can use the Gibbs sampler.

Markov chain

Markov Chain on the space Θ (state space) is a stochastic process satisfying the markov property

$$p(\theta^{(m+1)}|\theta^{(1)},\ldots,\theta^{(m)}) = p(\theta^{(m+1)}|\theta^{(m)})$$

The MC will explore the parameter space. The rule governing how to jump from one state to another is described with a transition kernel

Transition kernel

Consider a discrete state space of 3 states, i.e., θ can take 3 values. The corresponding transition matrix P is

$$\begin{pmatrix} p\left(\theta_A^{(m+1)}|\theta_A^{(m)}\right) & p\left(\theta_B^{(m+1)}|\theta_A^{(m)}\right) & p\left(\theta_C^{(m+1)}|\theta_A^{(m)}\right) \\ p\left(\theta_A^{(m+1)}|\theta_B^{(m)}\right) & p\left(\theta_B^{(m+1)}|\theta_B^{(m)}\right) & p\left(\theta_C^{(m+1)}|\theta_B^{(m)}\right) \\ p\left(\theta_A^{(m+1)}|\theta_C^{(m)}\right) & p\left(\theta_B^{(m+1)}|\theta_C^{(m)}\right) & p\left(\theta_C^{(m+1)}|\theta_C^{(m)}\right) \end{pmatrix}.$$

The rows sum to one and define a conditional probability mass function (conditional on the current state).

The columns are the marginal probabilities of being in a certain state in the next period.

This is naturally extended to continuous state spaces.

Stationary distribution

Let us denote as $\Pi^{(0)}$ the starting distribution (pmf)

at iteration m: $\Pi^{(m)}$ the distribution from which $\theta^{(m)}$ is drawn is

$$\Pi^{(m)} = \Pi^{(0)} \times P^m$$

We define the stationary distribution π to be some distribution such that $\pi = \pi P$.

our aim in Bayesian statistics generate a Markov chain whose stationary distribution is our posterior $\pi(\theta|x)$. From the random draws from the posterior we can use Monte Carlo principles to compute quantities of interest.

Difficulty: when has the chain converge? has it converged to the posterior dist. ?

Monte carlo principles for Markov chains

Beware: our draws are **not independent**, SLNN have been used to justify Monte Carlo Integration.

But we have an analog to SLLN for markov chains: the **Ergodic Theorem.**

Let $\left\{\theta^{(m)},\ 1 \leq m \leq M\right\}$ be M values from an aperiodic, irreducible, positive recurrent markov chain and $\mathbb{E}(g[\theta]) < \infty$, then

$$\frac{1}{M}\sum_{m=1}^{M} g(\theta^{(m)}) \to \int_{\Theta} g(\theta)\pi(\theta|x)d\theta,\ M \to \infty,$$

where π is the stationary distribution.

Gibbs sampler in a nutshell

The algorithm is:

- *step 0:* initialize $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_P^{(0)})$, set $m = 1$,
- *step 1:* for $p \in \{1, \ldots, P\}$ sample $\theta_p^{(m)}$ from $\pi(\theta_p | \theta_1^{(m)}, \ldots \theta_{p-1}^{(m)} \ldots \theta_{p+1}^{(m-1)}, \ldots, \theta_P^{(m-1)}, x)$
- *step 2:* set $m = m + 1$ and go back to step 1. Iterate until you obtain enough samples from the stationary distribution.

Reminder: Multivariate normal distribution

Let $X \sim \mathcal{N}_p(\mu, \Sigma)$, its pdf is given by:

$$f(x) = (2\pi)^{-p/2}|\Sigma|^{-\frac{1}{2}} \exp\left\{\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}, \ f : \mathbb{R}^p \to \mathbb{R}.$$

Mahalanobis transformation

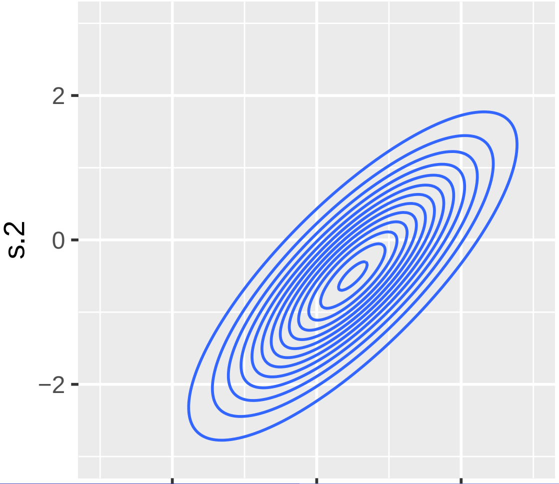
$$\begin{aligned} Y &= \Sigma^{-\frac{1}{2}}(X-\mu) \\ Y &\sim \mathcal{N}_p(0, I). \end{aligned}$$

Meaning that $Y_j \in \mathbb{R}$ the elements of Y are independent $\mathcal{N}(0,1)$. This implies that $f_Y(y) = \prod_{j=1}^{p} f_{Y_j}(y)$.

Geometry of the Multivariate Normal

The density of the $\mathcal{N}_p(\mu, \Sigma)$ forms ellipsoids of the form

$$(x-\mu)^t\Sigma^{-1}(x-\mu) = d^2.$$



Gibbs sampling for multivariate normal

Remark: on using properties of the multivariate normal for inverse probability inference.

Use the gibbs sampler to generate a sample of size 1000 from the joint distribution of (θ_1, θ_2) given by:

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

Draw a sample of size 1000 from this joint distribution.

Gibbs sampler for bivariate normal (solution)

From the properties of multivariate normal, we get:

$$\theta_1|\theta_2 \sim \mathcal{N}(\theta_2\rho, 1-\rho^2)$$
$$\theta_2|\theta_1 \sim \mathcal{N}(\theta_1\rho, 1-\rho^2)$$

Gibbs sampler for bivariate normal (solution)

```r
burn_in = 500
M = 10000 + burn_in
rho = 0.8
theta1= theta2 = rep(0, length = M)
theta1[1] = theta2[1] =10 # initial values

for (m in 2:M)
{
  theta1[m] = rnorm(1,mean = rho*theta2[m-1], sd = sqrt(1-rho^2))
  theta2[m] = rnorm(1,mean = rho*theta1[m], sd = sqrt(1-rho^2))
}
theta1 = theta1[-c(1:burn_in)] # burn-in
theta2 = theta2[-c(1:burn_in)] # burn-in
post = cbind(theta1,theta2)
colnames(post) = c("theta1","theta2")
```

Checking convergence

we expect convergence to a stationary distribution which is also our posterior

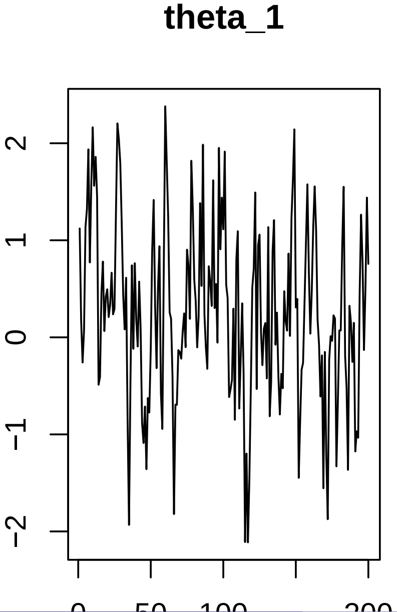
How to check?

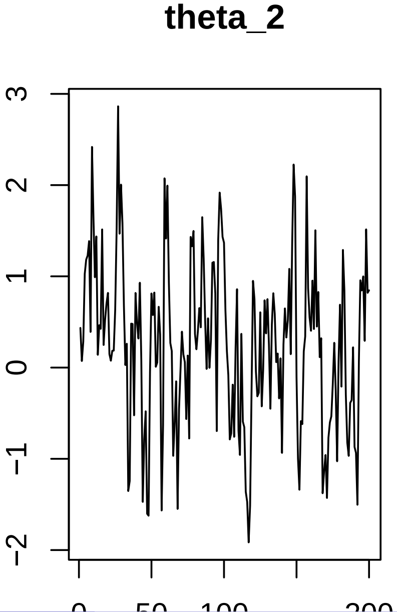
- **visual inspection:** how well chains are mixing
- **autocorrelation:** high autocorrelation = slow mixing
- **Rubin, Gelman, multiple chains diagnotic**

how to improve?

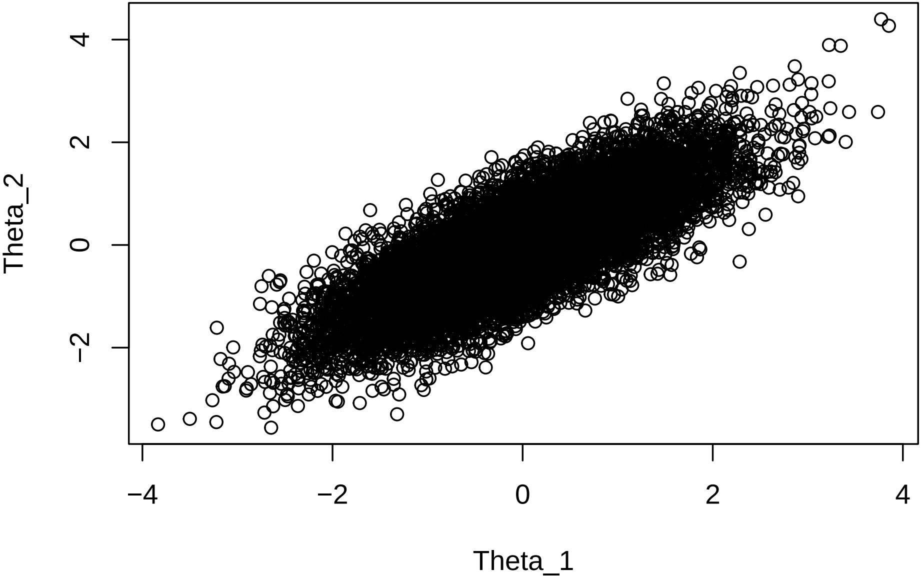
- **burnin:** discard first M generated values (till convergence of the chain to its stationary distribution)
- **thining:** keep every m-th observations in our chains to eliminate autocorrelation

Checking convergence





sample from the joint posterior



Gibbs sampler for Behrens-Fisher problem

Let us denote $D = (\bar{x}_1, \bar{x}_1, s_1^2, s_2^2)$.

The gibbs sampler can be written as:

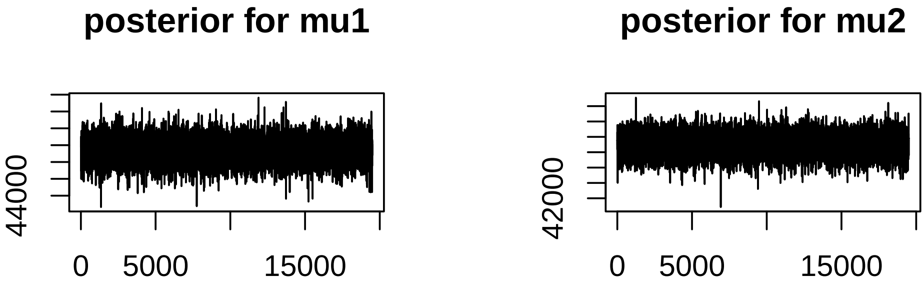
- step 0: initial values for $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$, set $m = 1$
- step 1:
  - $\mu_1^{(m)}$ draw from $\mu_1 | \mu_2^{(m-1)}, \sigma_1^{2,(m-1)}, \sigma_2^{2,(m-1)}, D \sim \mathcal{N}(\bar{x}_1, \sigma_1^{2,(m-1)}/n_1)$
  - $\mu_2^{(m)}$ draw from $\mu_2 | \mu_1^{(m)}, \sigma_1^{2,(m-1)}, \sigma_2^{2,(m-1)}, D \sim \mathcal{N}(\bar{x}_2, \sigma_2^{2,(m-1)}/n_2)$
  - $\sigma_1^{2,(m)}$ draw from $\sigma_1^2 | \mu_1^{(m)}, \mu_2^{(m)}, \sigma_2^{2,(m-1)}, D \sim \Gamma^{-1}\left(\frac{n_1}{2}, \frac{(n_1-1)s_1^2 + n_1\left(\bar{x}_1 - \mu_1^{(m)}\right)^2}{2}\right)$
  - $\sigma_2^{2,(m)}$ draw from $\sigma_2^2 | \mu_1^{(m)}, \mu_2^{(m)}, \sigma_1^{2,(m)}, D \sim \Gamma^{-1}\left(\frac{n_2}{2}, \frac{(n_2-1)s_2^2 + n_2\left(\bar{x}_2 - \mu_2^{(m)}\right)^2}{2}\right)$
- step 2: set $m \leftarrow m + 1$, iterate until $m = M$.

```
## Warning in .recacheSubclasses(def@className, def, env): undefined subcla
## "ndiMatrix" of class "replValueSp"; definition not updated

## ##
## ## Markov Chain Monte Carlo Package (MCMCpack)

## ## Copyright (C) 2003-2025 Andrew D. Martin, Kevin M. Quinn, and Jong He
```

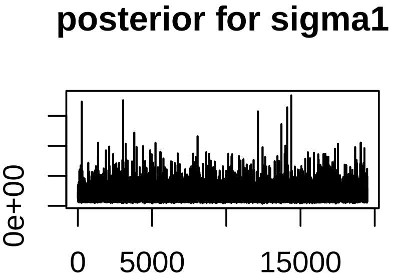
Gibbs sampler for Behrens-Fisher problem

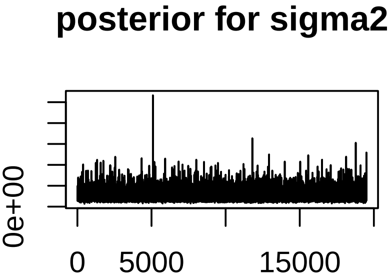
```r
x1.bar = mean(x1); x2.bar = mean(x2); s1 =  sd(x1); s2 =  sd(x2)
M= 20000;
mu1 = mu2 = sigma1 = sigma2 = rep(0,M)

# starting values
mu1[1] = x1.bar; mu2[1] = x2.bar; sigma1[1] = s1^2; sigma2[1] = s2^2

# iteration loop
for (m in 2:M)
{
  mu1[m] = rnorm(1, x1.bar, sqrt(sigma1[m-1]/n1))
  mu2[m] = rnorm(1, x2.bar, sqrt(sigma2[m-1]/n2))

  scale_val = 0.5*((n1-1)*s1^2+n1*(x1.bar-mu1[m])^2)
  sigma1[m] = rinvgamma(1,shape = n1/2, scale = scale_val)

  scale_val = 0.5*((n2-1)*s2^2+n2*(x2.bar-mu2[m])^2)
  sigma2[m] = rinvgamma(1,shape = n2/2, scale = scale_val)
}
```

Marginal posterior for the means

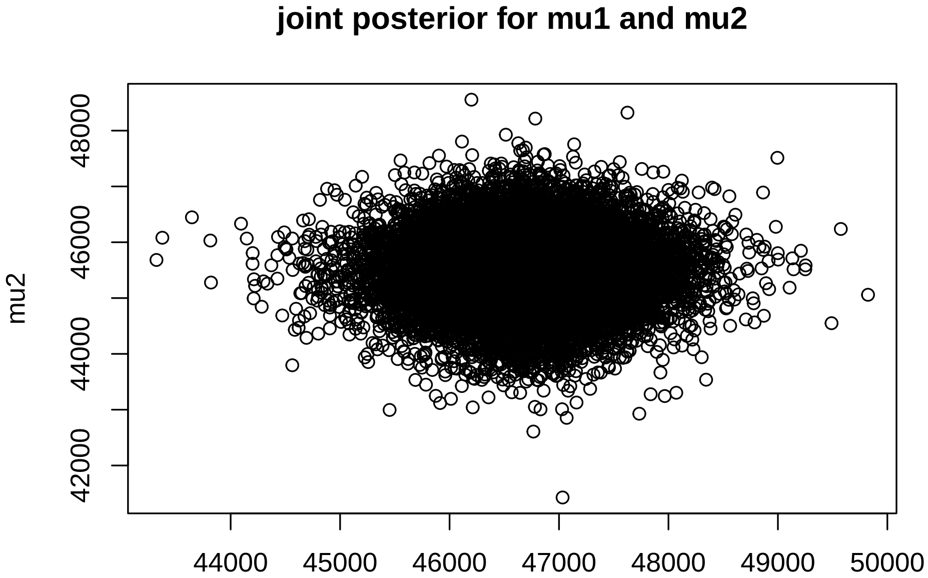


Marginal posterior for the variances





Joint posterior for the means



posterior inference for *delta*

```
delta= mu1-mu2
delta = delta[-c(1:burn_in)]
mean(delta); sd(delta)
```

```
## [1] 1193.113

## [1] 842.325
```

```
quantile(delta, c(0.025,0.05,0.5,0.95,0.975))
```

```
##      2.5%        5%       50%       95%     97.5%
## -467.4695 -189.3275 1188.9244 2581.7725 2859.6543
```

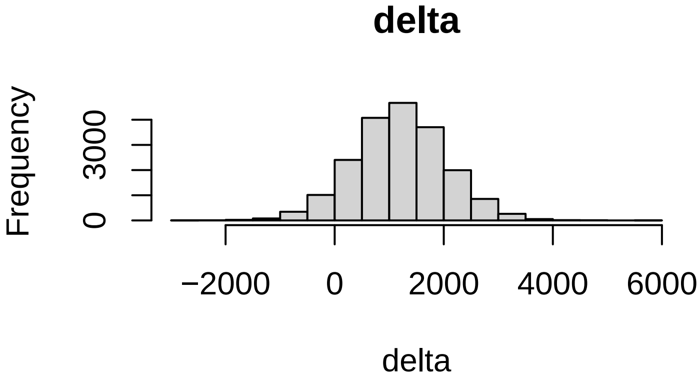
```
sum(delta>0)/length(delta)
```

```
## [1] 0.9248718
```

```
HPDinterval(as.mcmc(delta))
```

```
##          lower    upper
## var1 -454.5336 2870.633
## attr(,"Probability")
```

posterior inference for delta



Bayesian inference using R-Jags

```
library(coda, quietly = TRUE)
library(rjags, quietly = TRUE)
library(R2jags, quietly = TRUE)

model_code = '
model {
  for (i in 1:n1)
  {
    x1[i] ~ dnorm(mu1,tau1)
  }
  for (j in 1:n2)
  {
    x2[j] ~ dnorm(mu2,tau2)
  }
  mu1 ~ dnorm(0,0.0001)
  mu2 ~ dnorm(0,0.0001)
  tau1 <- 1/s1
  tau2 <- 1/s2
  delta <- mu1-mu2
  s1 ~ dgamma(0.0001, 0.0001)
  s2 ~ dgamma(0.0001, 0.0001)
```

Evolution of the Behrens-Fisher problem

We discussed the *initial* Behrens-Fisher problem and some approaches attemtping to tackle it.

There are also many settings in which one is facing *a type of* Behrens-Fisher problem in a case of:

- multivariate data:

$$x_{k1}, \dots, x_{kn_k} \sim \mathcal{N}_p(\mu_k, \Sigma_k), \ k = 1, 2.$$

Hypothesis testing problem:

$$H_0 : \ \mu_1 = \mu_2, \text{ vs } \mu_1 \neq \mu_2$$

Remark: if assume $\Sigma_1 = \Sigma_2$, then we have the Hotelling T^2-test.

- non-normal distributions (known or not)
- k-samples (generalized Behrens-Fisher problem)
- high-dimensional observations (number of observations $<<$ number of variables)
- in the context of **object Oriented Data Analysis** (OODA)

Object Oriented Data Analysis (OODA)

Big Data also means **more complex** data

OODA: *provides a framework for approaching complex data challenges (Marron & Wang, 2007).*

Important aspects of modern complex data analysis:

- what should be the *object* (most basic parts) of the statistical analysis?
- what should be the role of mathematics in the field ? $\rightarrow$ developing new methods

Following Marron and Wang, there are two main components of OODA:

- *3 phases:* object definition, exploratory analysis and confirmatory analysis
- *modes of variation*: a mode of variation of a sample of data objects is a set of potential members of the object space that provide a simple summary of one component of the variation.

Object Oriented Data Analysis (OODA)

Similarity with Oriented-object Programming? in the sense that **careful consideration of the data object tends to orient the analysis**.

object definition: determination of the data object **and** of its numerical representation. **each data object is thought as a point in a cloud of points**

object space: abstract space containing all possible objects

feature space: contains the practical numerical representations (e.g., data matrix) (numbers that male up a numerical representation).

It is important to keep in mind the nature of the object space during all the stages of the analysis.

Examples special topics of OODA: Circular data analysis, compositional data analysis, functional data analysis (FDA). . .

Evolution of the Behrens-Fisher problem

FDA: functional data analysis

- **data object** = functions/curves.
- **object space** = space of all functions or a subset (e.g., all monotonically increasing functions on $[0, T]$). It includes the choice of an appropriate metric. Typically in FDA, L^p family of norms, most often L^2 but for example in image analysis this norm is not the best choice according to visual human perception.
- **feature space** curves are represented as digitized vectors, the features are the entries of the vectors. Typical representation of functions using: Fourier, orthogonal polynomials, splines, wavelets. . . Hence the feature space though as space of vectors of the basis coefficients.

Evolution of the Behrens-Fisher problem

Behrens-Fisher: case of functional data

Consider two samples modelled as realizations of Gaussian processes $GP(\mu, \Sigma)$, where $\mu(t)$ is a mean function, $\Sigma(s,t)$ an autocovariance function, $s, t \in \mathcal{T}$ where $\mathcal{T}$ is a compact interval

$$y_{1,i}(t), \dots, y_{1,n_1}(t) \sim GP(\mu_1, \Sigma_1), \quad y_{2,j}(t), \dots, y_{1,n_2}(t) \sim GP(\mu_2, \Sigma_2),$$

Σ_1, Σ_2 are unknown and unequal. We then set the following hypothesis testing problem:

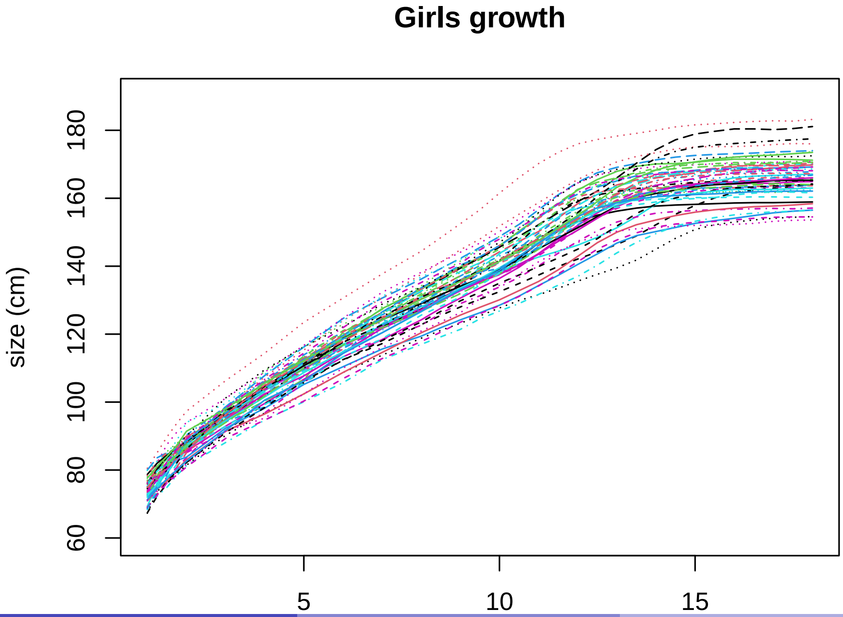
$$H_0 : \mu_1(t) = \mu_2(t) \text{ vs } H_1 : \mu_1(t) \neq \mu_2(t)$$

Evolution of the Behrens-Fisher problem

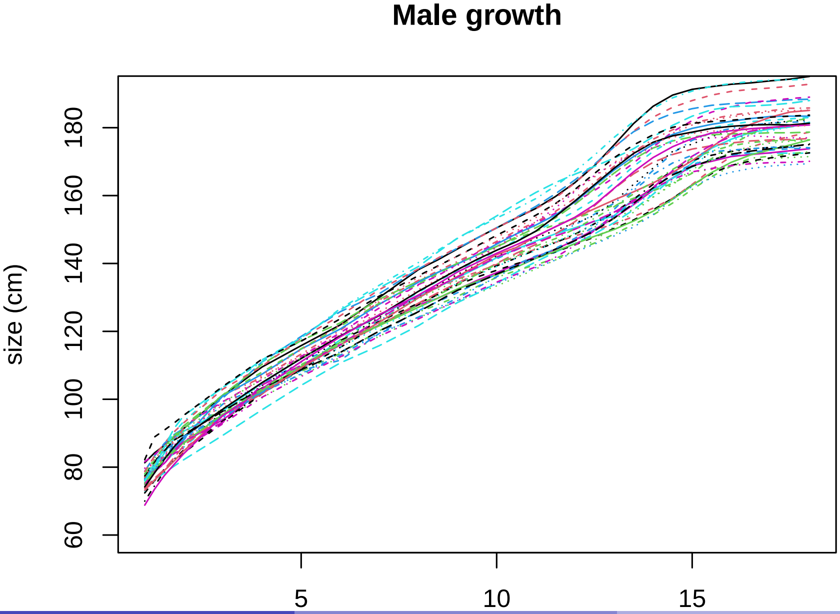
Data collected from Berkeley study on the growth of children and teenagers (Tuddenham and Snyder 1954).

- height of 54 girls and 39 boys measured at 31 ages from 1 year to 18 years old.
- observed ages are not equidistant
- these data from height are (should ?) considered as growth curves.

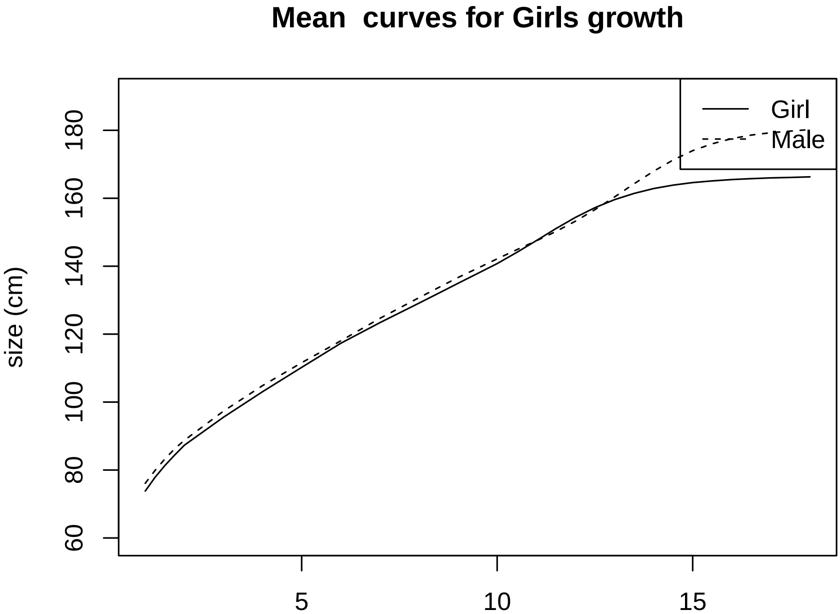
Evolution of the Behrens-Fisher problem



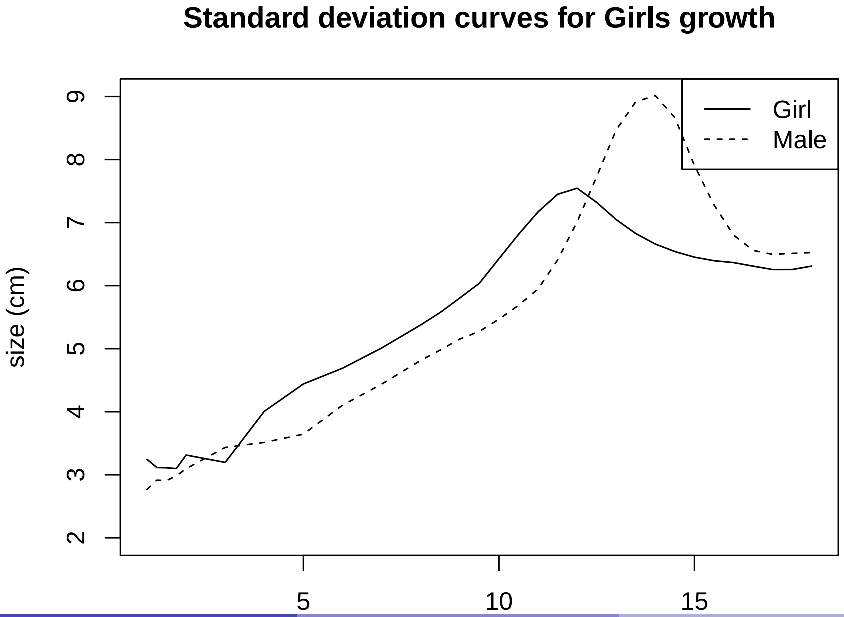
Evolution of the Behrens-Fisher problem



Evolution of the Behrens-Fisher problem



Evolution of the Behrens-Fisher problem



Evolution of the Behrens-Fisher problem

Questions:

- Do girls and boys grow at the same pace ?
- Do girls and boys grow at the same pace over different periods ?
  - baby (1-4 year)?
  - post-baby (4-13 year)?
  - teenage (13-18 year)?
- Graphs suggest that covariances are different between girls and boys.

Behrens-Fisher problem for functional data.

Main ingredient to find a good method: **find a good basis !**

Splines

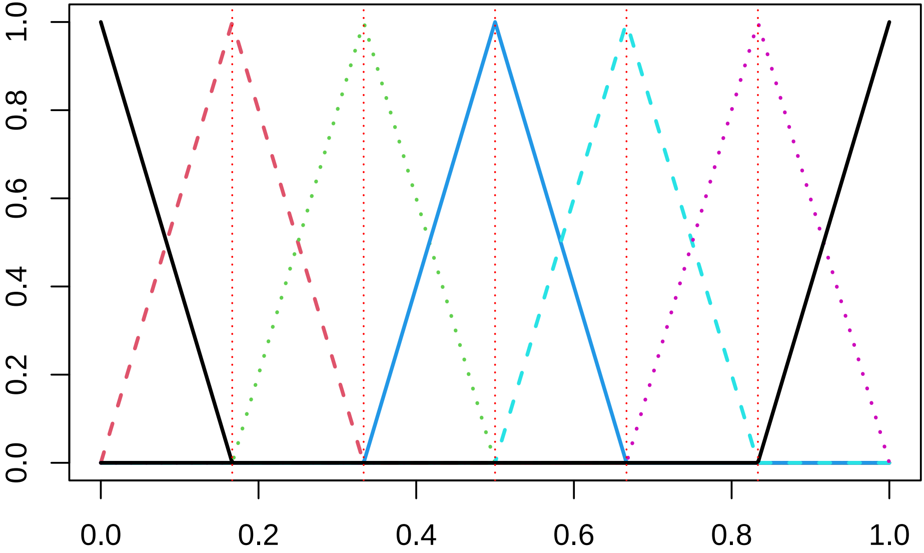
A spline on an interval $\mathcal{I} = [a, b]$ is a piecewise polynomial function with additional continuity conditions on the boundaries as well as for its derivatives. It is caracterized by its:

- order d : the maximal degree of polynomial on sub-intervals $+1$.
- knots that may not be equally spaced on $\mathcal{I}$.
- continuous derivatives up to the order $d - 2$.

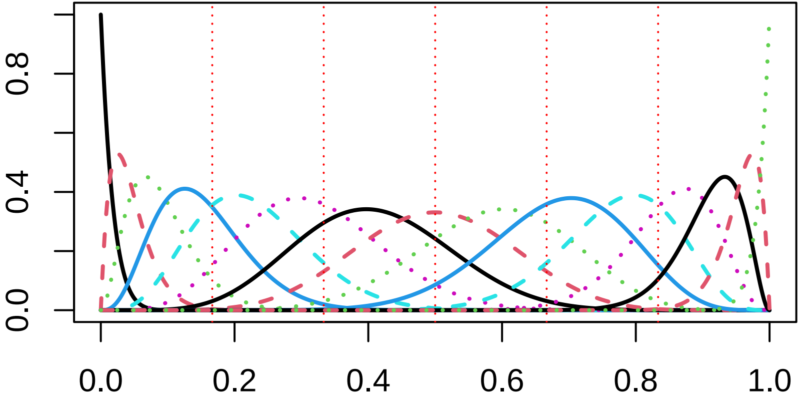
A spline basis of order d, associated to knots, is a family of functions such that:

- each basis function is a spline function.
- each spline of order d and knots can be expressed as a linear combination of these basis functions.
- the basis functions are linearly independent, not necessarily orthonormal.

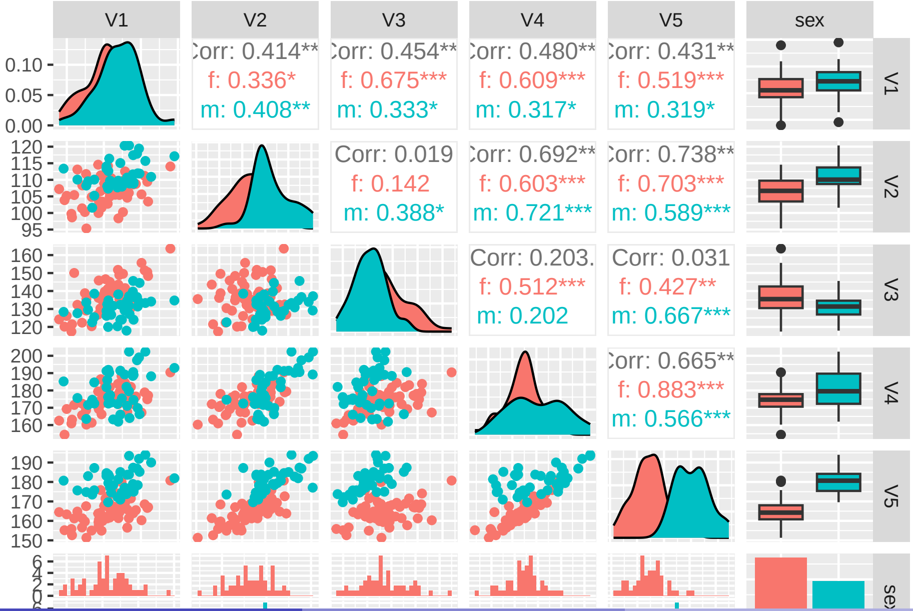
Example of spline basis



Example of spline basis



Splines coefficients for berkeley growth data



Evolution of the Behrens-Fisher problem

Jags two-sample test of multivariate normal means of spline coefficients.

```
## Inference for Bugs model at "4", fit using jags,
##  4 chains, each with 10000 iterations (first 200 discarded), n.thin = 2
##  n.sims = 19600 iterations saved
##              mu.vect sd.vect     2.5%      25%      50%      75%     97.5
## Sigma1[1,1]    9.920   1.984    6.790    8.525    9.683   11.031   14.53
## Sigma1[2,1]    4.727   2.061    1.113    3.314    4.569    5.965    9.19
## Sigma1[3,1]   20.625   5.154   12.300   16.943   20.014   23.591   32.39
## Sigma1[4,1]   13.905   3.766    7.741   11.277   13.457   16.076   22.59
## Sigma1[5,1]   10.255   3.116    5.033    8.071    9.920   12.067   17.33
## Sigma1[1,2]    4.727   2.061    1.113    3.314    4.569    5.965    9.19
## Sigma1[2,2]   19.819   3.931   13.566   17.067   19.330   22.023   29.01
## Sigma1[3,2]    6.157   6.094   -5.247    2.122    5.926    9.932   18.88
## Sigma1[4,2]   19.479   5.242   10.889   15.765   18.872   22.567   31.36
## Sigma1[5,2]   19.604   4.731   11.898   16.263   19.045   22.358   30.36
## Sigma1[1,3]   20.625   5.154   12.300   16.943   20.014   23.591   32.39
## Sigma1[2,3]    6.157   6.094   -5.247    2.122    5.926    9.932   18.88
## Sigma1[3,3]   94.078  18.554   64.584   81.002   91.857  104.797  136.66
## Sigma1[4,3]   36.046  11.085   17.492   28.304   34.836   42.733   60.54
## Sigma1[5,3]   25.994   9.258   10.202   19.583   25.085   31.560   46.21
```

Bibliography

- Droesbeke, J-J., Jeanne Fine, J., Saporta, G. (2002). Méthodes Bayésiennes en statistique. Technip (Paris).
- Ramsay, J.O, Giles Hooker, G., Spencer Graves, S. (2009). Functional Data Analysis with R and MATLAB. Springer.